LAWRENCE
LIVERMORE
NATIONAL
LABORATORY

UCRL-PRES-205398

Overview of the LSST Data Management Activities at LLNL

Ghaleb Abdulla

7/23/2004

Overview of the LSST Data Management Activities at LLNL

Ghaleb Abdulla

Center for Applied Scientific Computing

Lawrence Livermore National Lab

Overview

- Project, Team, collaborators
- Modular data-centric pipeline
- Support for real-time query processing
- H/W and S/W HPC architectures
- Other research topics
- Partnering approaches

Work in progress

LLNL/LSST Data Management

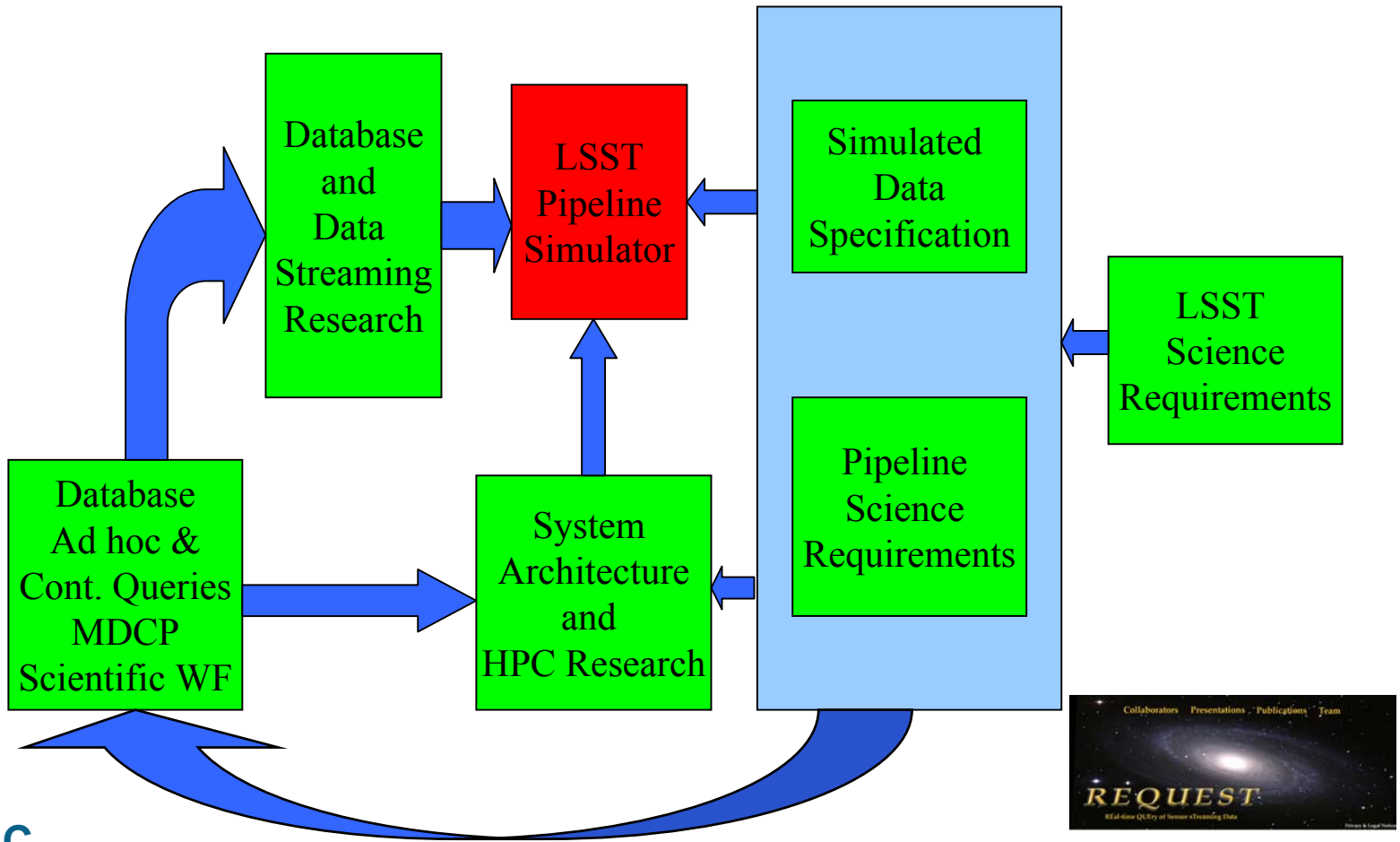
- **REal-time QUEry of Sensor sTreaming Data (REQUEST)**
  - http://www.llnl.gov/CASC/request/
- Team
  - Ghaleb Abdulla (100%,PL)
  - Kem Cook (20%, Project Scientist)
  - Sergei Nikolaev (50%, Application Scientist)
  - Marcus Miller (100%, System Architect)
  - Jim Garlick (50%, Advanced Developer)
  - Postdoc (100%, database/data-streaming-TBD)
- Collaborators
  - Berkeley
  - NCSA

LLNL Strengths

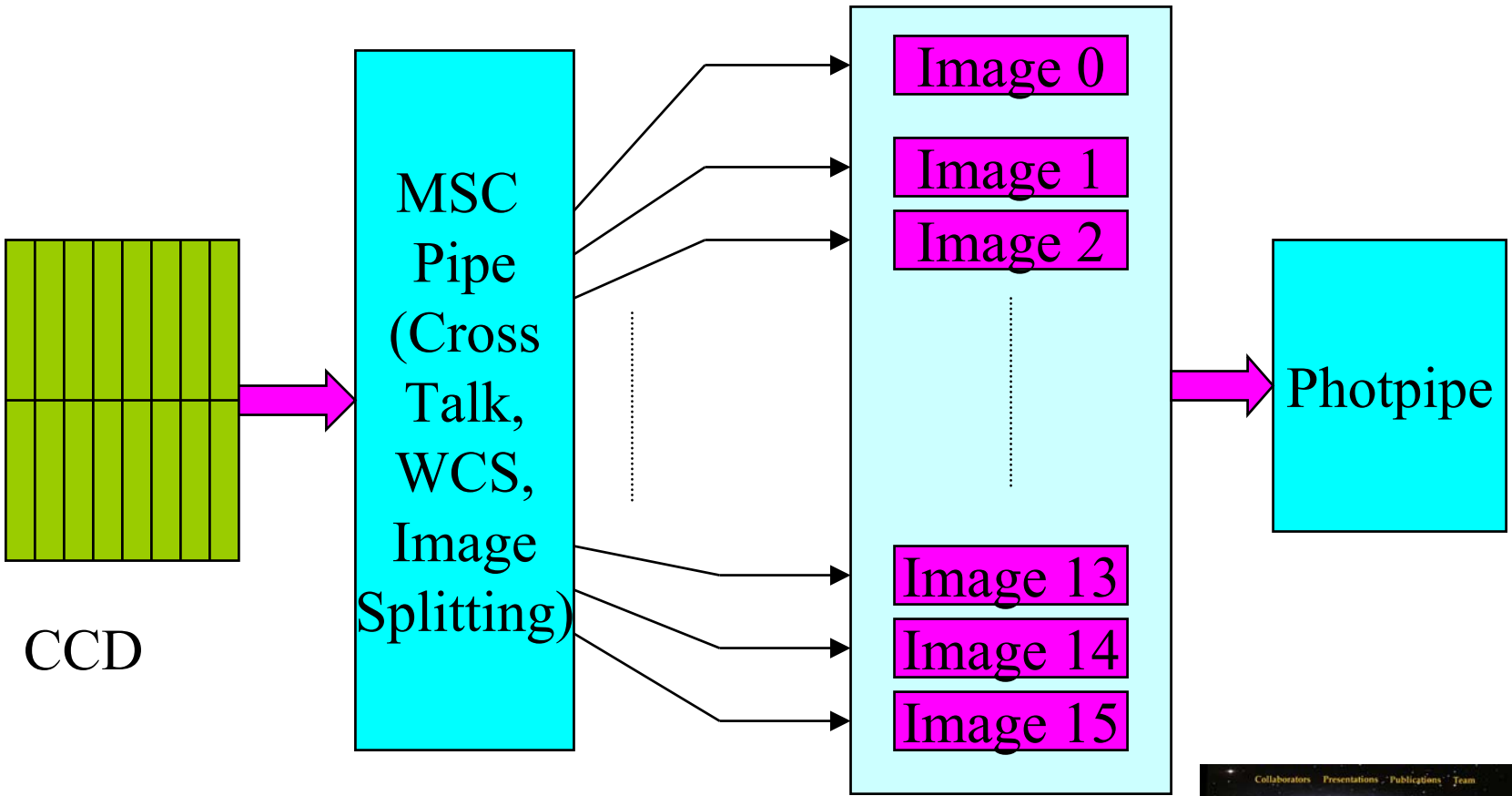
- Big cluster machines (MCR, Thunder, BlueGene/L, ….)
- HPC system architecture/development expertise
- Data-management research
- Opportunities to interact with the application scientists
- Strong potential collaborators (Berkeley, NCSA, …)

Overview Of LLNL Research Activities

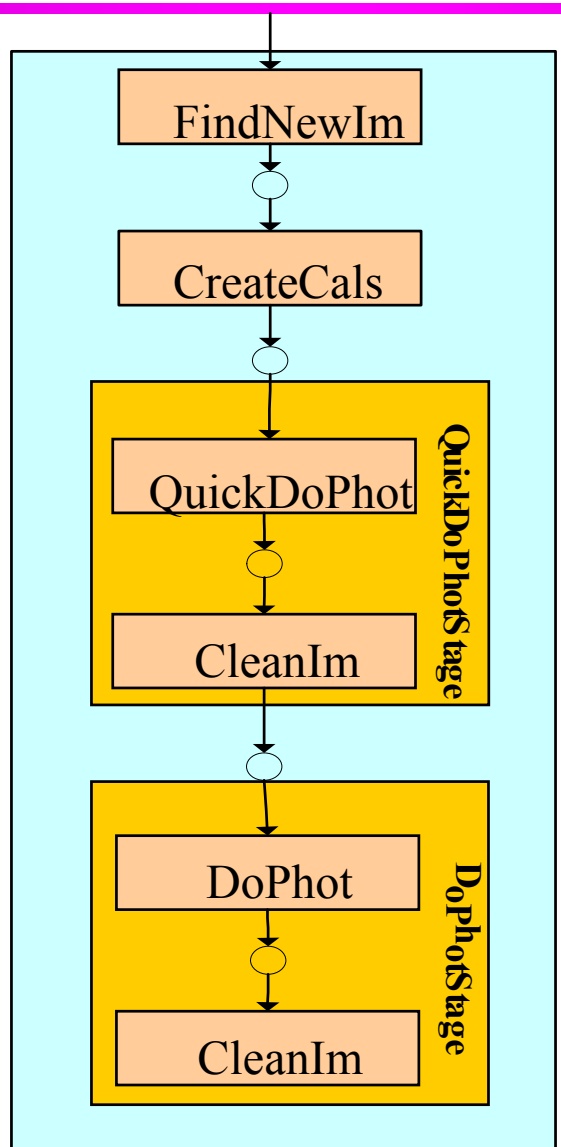
The objective is to build an LSST data pipeline simulator



SuperMacho Data Analysis Stages



Photpipe


FindNewIm
CreateCals
QuickDoPhot
CleanIm
QuickDoPhotStage
DoPhot
CleanIm
DoPhotStage


Collaborators
Presentations
Publications
Team
REQUEST

LLNL prototyping strategy and plans

- Use SM data pipeline to:
  - Test scalability on parallel cluster architectures
  - Research support of real time analysis
  - Generate synthetic data to test with anticipated LSST data rates
- Research thrusts to achieve these plans:
  - Modular data-centric pipeline
  - Data streaming and support of real-time continuous queries
  - H/W and S/W HPC architectures

Modular data-centric pipeline

- Collaborated with the database group at Berkeley (GridDB)
- Modeled photpipe, a super macho data pipeline, as one big box; implementation using GridDB is underway
- Started looking into modeling the individual pipeline components.

Milestones **(9/2004)**

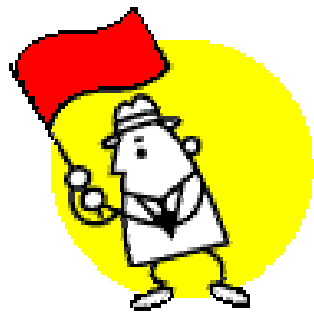
- **Show the utility of using a modular data centric software architectures by implementing the super macho data pipeline under GridDB**
- **Feed the output from the analysis into a database system**
- **Test the implementation with example queries**

Real time query and data analysis support

- Identify where real-time query support is needed within the data pipelines (Work in progress)
- Investigate using the data streaming system ,Telegraph, from Berkeley to support the real time query processing

Milestones (9/2004)

- **Identify a set of real-time continuous queries**
- **Investigate real-time support of the pre-defined set of queries for alerting services**
- **Investigate methods to integrate “Telegraph” with the the data pipeline work**



High Performance Computing

- Install and run the data pipelines on a cluster computer
- Characterize the data pipeline performance
- Investigate supporting scientific work-flows in a distributed computing environment

Milestones (9/2004)

- Port the existing Super Macho (SM) astrophysics image analysis pipeline to a large distributed memory, multi-node cluster platform employing a global parallel file system
- Evaluate the performance of the pipeline
- Profile the code to determine which sections yield the highest compute and I/O costs
- Characterize the common communication and I/O access patterns.

Milestones (9/2005)

- Generic implementation of GridDB that allows creating new analysis data pipelines with new functionalities by swapping algorithms and support parallel workflow scheduling on massively parallel clusters
- GUI to support query capabilities and easy workflow construction
- Integrate a streaming system to support real-time continuous queries with GridDB
  - Scalability issues
- Start tests of an LSST prototype pipeline

Other R&D issues

- LSST Digital Library
  - Real-time queries on digital libraries
  - Massive updates
  - Indexing issues (incremental, rebuilding the indexes)
- Confidence level and relevance in the final science results

Partnering Approaches
What We Have to Offer

- Technical contribution
  - Provide a framework for building new modular data centric analysis pipelines.
  - Help define requirements for flexible system architectures needed to support LSST science applications
  - Solve the technical research challenges that face the DM LSST efforts
- Informal leadership
  - Help setting up the direction of LSST DM efforts
  - Contribute in the grant review process
  - Help in other areas?

Partnering Approaches
Our Needs

- Money?
  - Enable academic collaborations
- Collaboration
  - Help us organize meetings and workshops
  - Provide us with feedback
  - Make our work known to other groups
- Include us in your plans

Questions